Data Scientist Role Play: Profiling and Analyzing the Yelp Dataset Coursera Worksheet

This is a 2-part assignment. In the first part, you are asked a series of questions that will help you profile and understand the data just like a data scientist would. For this first part of the assignment, you will be assessed both on the correctness of your findings, as well as the code you used to arrive at your answer. You will be graded on how easy your code is to read, so remember to use proper formatting and comments where necessary.

In the second part of the assignment, you are asked to come up with your own inferences and analysis of the data for a particular research question you want to answer. You will be required to prepare the dataset for the analysis you choose to do. As with the first part, you will be graded, in part, on how easy your code is to read, so use proper formatting and comments to illustrate and communicate your intent as required.

For both parts of this assignment, use this "worksheet." It provides all the questions you are being asked, and your job will be to transfer your answers and SQL coding where indicated into this worksheet so that your peers can review your work. You should be able to use any Text Editor (Windows Notepad, Apple TextEdit, Notepad ++, Sublime Text, etc.) to copy and paste your answers. If you are going to use Word or some other page layout application, just be careful to make sure your answers and code are lined appropriately.

In this case, you may want to save as a PDF to ensure your formatting remains intact for you reviewer.

Part 1: Yelp Dataset Profiling and Understanding

1. Profile the data by finding the total number of records for each of the tables below:

SELECT COUNT(*)

FROM [Table Name];

i. Attribute table = 10000

ii. Business table = 10000

iii. Category table =10000

iv. Checkin table =10000

v. elite_years table =10000

vi. friend table = 10000

vii. hours table =10000

viii. photo table = 10000

ix. review table = 10000

x. tip table = 10000

xi. user table =10000

2. Find the total distinct records by either the foreign key or primary key for each table. If two foreign keys are listed in the table, please specify which foreign key.

```
SELECT COUNT(distinct([key]))
FROM [table name];
```

i. Business = 10000

ii. Hours = 1562

iii. Category = 2643

iv. Attribute = 1115

v. Review = 10000

vi. Checkin = 493

vii. Photo = 10000

viii. Tip = 3979

ix. User = 10000

x. Friend = 11

xi. Elite_years = 2780

Note: Primary Keys are denoted in the ER-Diagram with a yellow key icon.

3. Are there any columns with null values in the Users table? Indicate "yes," or "no."

Answer: no

SQL code used to arrive at answer:

```
Select Count(*)
FROM user
WHERE id IS NULL OR
name IS NULL OR
review_count IS NULL OR
yelping_since IS NULL OR
useful IS NULL OR
funny IS NULL OR
cool IS NULL OR
fans IS NULL OR
average_stars IS NULL OR
compliment_hot IS NULL OR
compliment_more IS NULL OR
compliment_profile IS NULL OR
compliment_cute IS NULL OR
compliment_list IS NULL OR
compliment_note IS NULL OR
compliment_plain IS NULL OR
compliment_cool IS NULL OR
compliment_funny IS NULL OR
compliment_writer IS NULL OR
compliment_photos IS NULL
```

4. For each table and column listed below, display the smallest (minimum), largest (maximum), and average (mean) value for the following fields:

```
select Min(review_count), Max(review_count), Avg(review_count)
from user;
```

i. Table: Review, Column: Stars

min:1 max:5 avg:3.7082

ii. Table: Business, Column: Stars

min:1 max:5 avg:3.6549

iii. Table: Tip, Column: Likes

min:0 max:2 avg:0.0144

iv. Table: Checkin, Column: Count

min:1 max:53 avg:1.9414

v. Table: User, Column: Review_count

min:0 max:2000 avg:24.2995

5. List the cities with the most reviews in descending order:

SQL code used to arrive at answer:

```
select city, review_count
from business
order by review_count desc;
```

Copy and Paste the Result Below:

```
+------------------+---------+
| city             | total_r |
+------------------+---------+
| Las Vegas        |   82854 |
| Phoenix          |   34503 |
| Toronto          |   24113 |
| Scottsdale       |   20614 |
| Charlotte        |   12523 |
| Henderson        |   10871 |
| Tempe            |   10504 |
| Pittsburgh       |    9798 |
| Montréal         |    9448 |
| Chandler         |    8112 |
| Mesa             |    6875 |
| Gilbert          |    6380 |
| Cleveland        |    5593 |
| Madison          |    5265 |
| Glendale         |    4406 |
| Mississauga      |    3814 |
| Edinburgh        |    2792 |
| Peoria           |    2624 |
| North Las Vegas  |    2438 |
| Markham          |    2352 |
| Champaign        |    2029 |
| Stuttgart        |    1849 |
| Surprise         |    1520 |
| Lakewood         |    1465 |
| Goodyear         |    1155 |
+------------------+---------+
(Output limit exceeded, 25 of 362 total rows shown)
```

6. Find the distribution of star ratings to the business in the following cities:

i. Avon

SQL code used to arrive at answer:

```
select city, count(stars)
from business
where city = 'Avon'
group by stars
```

Copy and Paste the Resulting Table Below (2 columns â€“ star rating and count):

```
+------+--------------+
| city | count(stars) |
+------+--------------+
| Avon |            1 |
| Avon |            2 |
| Avon |            3 |
| Avon |            2 |
| Avon |            1 |
| Avon |            1 |
+------+--------------+
```

ii. Beachwood

SQL code used to arrive at answer:

```
select city, count(stars)
from business
where city = 'Beachwood'
group by stars
```

Copy and Paste the Resulting Table Below (2 columns â€“ star rating and count):

```
+-----------+--------------+
| city      | count(stars) |
+-----------+--------------+
| Beachwood |            1 |
| Beachwood |            1 |
| Beachwood |            2 |
| Beachwood |            2 |
| Beachwood |            1 |
| Beachwood |            2 |
| Beachwood |            5 |
+-----------+--------------+
```

7. Find the top 3 users based on their total number of reviews:

SQL code used to arrive at answer:

```
select name, review_count
from user
order by review_count desc
limit 3
```

Copy and Paste the Result Below:

```
+--------+--------------+
| name   | review_count |
+--------+--------------+
| Gerald |         2000 |
| Sara   |         1629 |
| Yuri   |         1339 |
+--------+--------------+
```

8. Does posing more reviews correlate with more fans?

Please explain your findings and interpretation of the results: No, as the table below illustrates, there is no strict correlation between reviews and fans.

```
+-----------+--------------+------+
| name      | review_count | fans |
+-----------+--------------+------+
| Amy       |          609 |  503 |
| Mimi      |          968 |  497 |
| Harald    |         1153 |  311 |
| Gerald    |         2000 |  253 |
| Christine |          930 |  173 |
| Lisa      |          813 |  159 |
| Cat       |          377 |  133 |
| William   |         1215 |  126 |
| Fran      |          862 |  124 |
| Lissa     |          834 |  120 |
| Mark      |          861 |  115 |
| Tiffany   |          408 |  111 |
| bernice   |          255 |  105 |
| Roanna    |         1039 |  104 |
| Angela    |          694 |  101 |
| .Hon      |         1246 |  101 |
| Ben       |          307 |   96 |
| Linda     |          584 |   89 |
| Christina |          842 |   85 |
| Jessica   |          220 |   84 |
| Greg      |          408 |   81 |
| Nieves    |          178 |   80 |
| Sui       |          754 |   78 |
| Yuri      |         1339 |   76 |
| Nicole    |          161 |   73 |
+-----------+--------------+------+
```

9. Are there more reviews with the word "love" or with the word "hate" in them?

Answer: The number of reviews containing love is 1780 and hate is 232

SQL code used to arrive at answer:

```
select count(id) as number_of_neg_reviews
from review
where text like '%hate%';

select count(id) as number_of_pos_reviews
from review
where text like '%love%';
```

10. Find the top 10 users with the most fans:

SQL code used to arrive at answer:

```
select name, fans
from user
order by fans desc
limit 10;
```

Copy and Paste the Result Below:

```
+-----------+------+
| name      | fans |
+-----------+------+
| Amy       |  503 |
| Mimi      |  497 |
| Harald    |  311 |
| Gerald    |  253 |
| Christine |  173 |
| Lisa      |  159 |
| Cat       |  133 |
| William   |  126 |
| Fran      |  124 |
| Lissa     |  120 |
+-----------+------+
```

Part 2: Inferences and Analysis

1. Pick one city and category of your choice and group the businesses in that city or category by their overall star rating. Compare the businesses with 2-3 stars to the businesses with 4-5 stars and answer the following questions. Include your code.

i. Do the two groups you chose to analyze have a different distribution of hours?

Yes

ii. Do the two groups you chose to analyze have a different number of reviews?

Yes, 4-5 stars group has more reviews (32).

iii. Are you able to infer anything from the location data provided between these two groups? Explain.

Higher rated companies have no neighborhood info and all of them are on Tropicana Ave.

SQL code used for analysis:

```
SELECT
    CASE
        WHEN is_open THEN 'Open'
        ELSE 'Closed'
        END as Availability,
       review_count,
       hours.hours,
       name,
       address,
       neighborhood

FROM business b INNER JOIN category c on b.id = c.business_id
    INNER JOIN hours on b.id = hours.business_id
WHERE city = 'Las Vegas' AND category = 'Shopping' AND (stars>=4 OR (stars<4 AND stars>=2))
ORDER BY stars DESC, hours DESC
```

2. Group business based on the ones that are open and the ones that are closed. What differences can you find between the ones that are still open and the ones that are closed? List at least two differences and the SQL code you used to arrive at your answer.

i. Difference 1: The number of open businesses is much higher.

ii. Difference 2: Open businesses have more reviews.

SQL code used for analysis:

```sql
select is_open, count(id) as num_of_bussiness, avg(stars), avg(review_count)
from business
group by is_open;
```

3. For this last part of your analysis, you are going to choose the type of analysis you want to conduct on the Yelp dataset and are going to prepare the data for analysis.

Ideas for analysis include: Parsing out keywords and business attributes for sentiment analysis, clustering businesses to find commonalities or anomalies between them, predicting the overall star rating for a business, predicting the number of fans a user will have, and so on. These are just a few examples to get you started, so feel free to be creative and come up with your own problem you want to solve. Provide answers, in-line, to all of the following:

i. Indicate the type of analysis you chose to do:

I want to get insight into business type and city correlation for an entrepreneur who wants to establish a company in a certain city.

ii. Write 1-2 brief paragraphs on the type of data you will need for your analysis and why you chose that data:

I need city, category, stars, and review_count data. By grouping businesses in Las Vegas according to their category, we can get better insight into every business type. I used review count to estimate user number/business. And, dividing review count into stars gives information about easiness of making profit without pretentious service or product.

## iii. Output of your finished dataset:

```
+-----------+------------------------------+-----------------+--------------+--------------+--------------+--------------+--------------------------------+
| city      | category                     | easy_profit_idx | avg(b.stars) | review_count | max(b.stars) | min(b.stars) | name                           |
+-----------+------------------------------+-----------------+--------------+--------------+--------------+--------------+--------------------------------+
| Las Vegas | Malaysian                    |           192.0 |          4.0 |          768 |          4.0 |          4.0 | Big Wong Restaurant            |
| Las Vegas | Taiwanese                    |           192.0 |          4.0 |          768 |          4.0 |          4.0 | Big Wong Restaurant            |
| Las Vegas | French                       |            42.0 |          4.0 |          168 |          4.0 |          4.0 | Jacques Cafe                   |
| Las Vegas | Vegetarian                   |            42.0 |          4.0 |          168 |          4.0 |          4.0 | Jacques Cafe                   |
| Las Vegas | Arcades                      |           26.25 |          4.0 |          105 |          4.5 |          3.5 | Hi Scores - Blue Diamond       |
| Las Vegas | Chocolatiers & Shops         |             7.5 |          4.0 |           30 |          4.0 |          4.0 | Sweet Ruby Jane Confections    |
| Las Vegas | Community Service/Non-Profit |   7.11111111111 |          4.5 |           32 |          4.5 |          4.5 | Red Rock Canyon Visitor Center |
| Las Vegas | Special Education            |   7.11111111111 |          4.5 |           32 |          4.5 |          4.5 | Red Rock Canyon Visitor Center |
| Las Vegas | Visitor Centers              |   7.11111111111 |          4.5 |           32 |          4.5 |          4.5 | Red Rock Canyon Visitor Center |
| Las Vegas | Education                    |   6.73684210526 |         4.75 |           32 |          5.0 |          4.5 | Red Rock Canyon Visitor Center |
+-----------+------------------------------+-----------------+--------------+--------------+--------------+--------------+--------------------------------+
```

## iv. Provide the SQL code you used to create your final dataset:

```
select city, category, (review_count/avg(stars)) as easy_profit_idx, avg(stars), review_count, max(stars),min(stars),name

from business b inner join category c on b.id=c.business_id

group by category

having city = 'Las Vegas'

order by easy_profit_idx desc

limit 10
```